THAT WHICH IS CLAIMED IS:

1. A method of distributing workload between a plurality of servers, the method comprising:

receiving a plurality of requests over a first connection;

parsing the plurality of requests to determine application layer information associated with each of the plurality of requests;

selecting destination servers for corresponding ones of the plurality of requests based on the determined application layer information associated with each of the plurality of requests; and

distributing the plurality of requests to the corresponding selected destination servers over a plurality of second connections associated with respective ones of the destination servers.

2. A method according to Claim 1, wherein the first connection comprises an HTTP 1.1 connection.

3. A method according to Claim 1, wherein parsing the plurality of requests comprises:

determining a start point and an end point for each of the plurality of requests within the first connection; and

identifying application layer information within each of the plurality of requests.

4. A method according to Claim 3, wherein the application layer information comprises layer 7 information and above.

5. A method according to Claim 3, wherein the application layer information comprises at least one of a type of request, a client identification, an individual user identification, and a cookie.

6. A method of Claim 1, wherein the plurality of requests comprise a plurality of Hypertext Transport Protocol(HTTP) requests.

7. A method according to Claim 1, wherein selecting destination servers for corresponding ones of the plurality of requests comprises:

determining if the determined application layer information associated with each of the plurality of requests is relevant application layer information;

selecting one of a subset of the destination servers if the application layer information associated with each of the plurality of requests is relevant application layer information; and

selecting a destination server other than a destination server in the subset of the destination servers if the application layer information associated with each of the plurality of requests is not relevant application layer information.

8. A method of Claim 7, wherein selecting one of a subset of the destination servers if the application layer information associated with each of the plurality of requests is relevant application layer information, further comprises:

determining a load associated with respective destination servers in the subset of destination servers; and

selecting the destination server in the subset of the destination servers based on the determined load.

9. A method of Claim 7, wherein the subset of destination servers includes at least one server which is to receive requests having an indication of high priority, and wherein the indication of high priority is determined based on the existence and nonexistence of relevant application layer information.

10. A method according to Claim 1, wherein distributing the plurality of requests comprises:

determining if a second connection associated with a selected destination servers exists;

establishing the second connection to the selected destination server if the second connection does not exist;

distributing a request to the selected destination servers over the second connection; and

repeating the determining, establishing and distributing for each of the plurality of requests.

11. A method according to Claim 1, wherein receiving, parsing, selecting and distributing are carried out by an application executing on a data processing system.

12. A method according to Claim 1, further comprising tracking the plurality of requests and a plurality of corresponding responses to the plurality of requests.

13. A method according to Claim 1, wherein distributing the plurality of requests, comprises:

routing the plurality of requests using network address translation at a routing layer of a communication protocol stack.

14. A method according to Claim 13, wherein routing the plurality of requests further comprises routing the plurality of requests using session control translation at the routing layer of the communication protocol stack.

15. A method according to Claim 14, wherein routing the plurality of requests includes routing the corresponding responses to the plurality of requests using network address translation at a routing layer of a communication protocol stack.

16. A method of distributing workload between a plurality of servers, wherein each of the plurality of servers is executing an instance of an application which communicates over a network such that each of a plurality of HTTP requests within a single HTTP 1.1 connection to the application may be distributed to any one of the plurality of servers, the method comprising:

defining a subset of the plurality of servers which are to receive HTTP requests having an indication of high priority;

establishing an HTTP 1.1 connection responsive to receiving a request for an HTTP 1.1 connection to the application over the network;

receiving a first Hypertext Transport Protocol(HTTP) request within the HTTP 1.1 connection;

parsing the first HTTP request to determine if the first HTTP request has an indication of high priority based on application layer information included in the first HTTP request; and

distributing the first HTTP request to one of the subset of the plurality of servers over a first connection if the first HTTP request has an indication of high priority.

17. A method according to Claim 16, further comprising:

distributing the first HTTP request to a server other than a server in the subset of the destination servers if the first HTTP request does not have an indication of high priority.

18. The method according to Claim 16, further comprising:

receiving a second HTTP request within the HTTP 1.1 connection

parsing the second HTTP request to determine if the second HTTP request has an indication of high priority based on application layer information included in the second HTTP request;

distributing the second HTTP request to one of the subset of the plurality of servers over a second connection if the second HTTP request has an indication of high priority; and

repeating the receiving, parsing and distributing steps for each subsequent HTTP request received within the HTTP 1.1 connection.

19. A method according to Claim 16, wherein distributing the first HTTP request, further comprises:

determining a load associated with respective servers in the subset of the plurality of servers; and

distributing the first HTTP request to the server in the subset of the plurality of servers based on the determined load.

20. A method according to Claim 16, wherein the indication of high priority is based on the existence and nonexistence of relevant application layer information.

21. A method according to Claim 20, wherein the application layer information comprises at least one of a type of request, a client identification, an individual user identification, and a cookie.

22. A method according to Claim 20, wherein the application layer information comprises layer 7 information and above.

23. A method according to Claim 16, wherein parsing the first HTTP request comprises:

determining a start point and an end point for the first HTTP request within the HTTP 1.1 connection;

identifying application layer information within the first HTTP request; and

determining if the application layer information is relevant application layer information.

24. A method according to Claim 16, wherein distributing the first HTTP request comprises:

determining if a first connection exists;

establishing the first connection if the first connection does not exist; and

distributing the first HTTP request over the first connection.

25. A method according to Claim 16, wherein the steps of defining, receiving, parsing, and distributing are carried out by an application executing on a data processing system.

26. A method according to Claim 16, further comprising tracking the HTTP request and a corresponding response to the HTTP request.

27. A method according to Claim 16, wherein distributing the first HTTP request comprises:

routing the first HTTP request using network address translation at a routing layer at a communication protocol stack.

28. A method according to Claim 27, wherein routing the first HTTP request further comprises routing the first HTTP request using session control translation at the routing layer at the communication protocol stack.

29. A method according to Claim 28, wherein routing the first HTTP request includes routing the corresponding response to the first HTTP request using network address translation at a routing layer of a communication protocol stack.

30. A system for distributing workload between a plurality of servers, comprising:

means for receiving a plurality of requests over a first connection;

means for parsing the plurality of requests to determine application layer information associated with each of the plurality of requests;

means for selecting destination servers for corresponding ones of the plurality of requests based on the determined application layer information associated with each of the plurality of requests; and

means for distributing the plurality of requests to the corresponding selected destination servers over a plurality of second connections associated with respective ones of the destination servers.

31. A system for distributing workload between a plurality of servers, wherein each of the plurality of servers is executing an instance of an application which communicates over a network such that each of a plurality of HTTP requests within a single HTTP 1.1 connection to the application may be distributed to any one of the plurality of servers, comprising:

means for defining a subset of the plurality of servers which are to receive HTTP requests having an indication of high priority;

means for establishing an HTTP 1.1 connection responsive to receiving a request for an HTTP 1.1 connection to the application over the network;

means for receiving a first Hypertext Transport Protocol(HTTP) request within the HTTP 1.1 connection;

means for parsing the first HTTP request to determine if the first HTTP request has an indication of high priority based on application layer information included in the first HTTP request; and

means for distributing the first HTTP request to one of the subset of the plurality of servers over a first connection if the first HTTP request has an indication of high priority.

32. A computer program product for distributing workload between a plurality of servers, comprising:

a computer readable program medium having computer readable program code embodied therein, the computer readable program code comprising:

computer readable program code which receives a plurality of requests over a first connection;

computer readable program code which parses the plurality of requests to determine application layer information associated with each of the plurality of requests;

computer readable program code which selects destination servers for corresponding ones of the plurality of requests based on the determined application layer information associated with each of the plurality of requests; and

computer readable program code which distributes the plurality of requests to the corresponding selected destination servers over a plurality of second connections associated with respective ones of the destination servers.

33. A computer program product for distributing workload between a plurality of servers, wherein each of the plurality of servers is executing an instance of an application which communicates over a network such that each of a plurality of HTTP requests within a single HTTP 1.1 connection to the application may be distributed to any one of the plurality of servers, comprising:

a computer readable program medium having computer readable program code embodied therein, the computer readable program code comprising:

computer readable program code which defines a subset of the plurality of servers which are to receive HTTP requests having an indication of high priority;

computer readable program code which establishes an HTTP 1.1 connection responsive to receiving a request for an HTTP 1.1 connection to the application over the network;

computer readable program code which receives a first Hypertext Transport Protocol(HTTP) request within the HTTP 1.1 connection;

computer readable program code which parses the first HTTP request to determine if the first HTTP request has an indication of high priority based on application layer information included in the first HTTP request; and

computer readable program code which distributes the first HTTP request to one of the subset of the plurality of servers over a first connection if the first HTTP request has an indication of high priority.